木に基づくクラスタリングを利用した音節波形接続型音声合成法*

◎植村和久, 村上仁一, 池原悟 (鳥取大)

1　はじめに

音声合成法の手法の 1 つとして，音節波形接続型音声合成法 [1] が提案されている．この手法は，録音した音声波形の一部（以下，音節素片）を取り出し，接続することによって合成音声を作成する．音声波形に信号処理を加えないため，自然性の高い音声が作成出来るが，音節素片選択時に，全ての条件を一致させなければならない．そのため，この手法の問題点の 1 つとして，任意の一般名詞を作成する際に大量の録音単語が必要となる．

その問題を解決するために，木に基づくクラスタリング（以下，クラスタリング）を利用した手法が提案されている [2]．この手法では， 音節素片選択時の全ての条件を完全に一致させるのではなく，一部の条件をクラスタリングを用いて緩和することにより，理論上全ての合成音声が作成可能となる．しかし，音節素片選択時の条件を緩和したことにより，音声品質が非常に悪い音声もできる．またデータベース中に存在しない単語に対する本手法の有効性もまだ示されていない．

そこで本研究では，データベース中に存在する音節素片に対して，直接クラスタリングを行う標準モデルと，アクセントを考慮しクラスタリングを行う拡張モデルを用いて合成音声を作成し，音声品質の調査を行う．またデータベース中に存在しない音節素片に対しても同様に両モデルを用いて合成音声を作成し，音声品質の調査を行う．

2　音節波形接続型音声合成法

2.1　音節波形接続型音声合成の概要

音節波形接続型音声合成 [1] では，音響的なパラメータを使用せずに Table 1 に示す言語的な情報を用いて合成音声を作成する．まず，データベース中の各音節素片に対して，単語のモーラ数，モーラ位置，アクセント型，各モーラ位置のアクセントの高低，さらに前後の音素環境のラベルを付与する．

次に合成する単語内に含まれる音節素片と，言語的な情報が一致する音節素片を選択する．最後に選択した音節素片を接続して合成音声を作成する．

例えば「乗り物 (no/ri/mo/no)」という合成音声を作成する際の例を Fig. 1 に示す．なお Fig. 1 の「_ | ¯」はアクセントの高低を表す．また，太文字で示している部分は，接続する音節素片である．

Table 1　音節波形接続型音声合成における言語的な情報

1. 中心の音節
2. 直前の音素 (前音素環境)
3. 直後の音素 (後音素環境)
4. 単語のモーラ数
5. 単語のモーラ位置
6. 単語のアクセント型
7. 単語のアクセントの高低

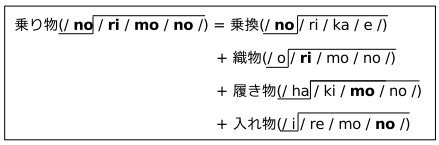

Fig. 1　音節波形接続型音声合成法の例

3　クラスタリング

3.1　クラスタリングの概要

クラスタリング [2] は，決定木に基づいて，質問を用いることで，データベース中に存在する音節素片を音響的特徴の類似した音節素片同士の状態集合（以後，クラスタ）に分類する．任意の音節素片も同様に各クラスタに分類できる．その際同じクラスタに含まれる音節素片を利用することで，任意の単語が作成可能となる．本研究では，Table 1 で示した言語的な情報に関する質問を作成し，HTK[3] の HHEd を用いることでクラスタリングを行う．

3.2　クラスタリングを利用した音節波形接続型音声合成法

本研究では，まずデータベース中の音節素片に対してクラスタリングを行う．次に合成したい単語の音節素片を含むクラスタの中から，音節素片を 1 つ抽出する．その後，音節波形接続型音声合成法で合成音声を作成する．

4　標準モデルと拡張モデル

本研究では，データベース中の音節素片全てに対して直接クラスタリングを行う標準モデルと，事前に音節素片全てをアクセント型（6 種類）およびアクセントの高低（2 種類）で分類した後にクラスタリングを行う拡張モデルを使用する．各モデルの詳細を Fig. 2 および 4 に示す．また各モデルを用いて Fig. 1 と同じ音声を作成した際の例をそれぞれ Fig. 3 および 5 に示す．Fig. 1 と比較すると，合成に使用した音節素片全てが，モーラ数やアクセント型の異なる音節素片であることが分かる．

また Fig. 4 中※で記述してある 3 桁の数字は，最初の 2 桁がアクセント型を示し，最後の 1 桁がアクセントの高低を示している．例えば “000” の場合，アクセント型が 00 型，アクセントの高低が 0 である．

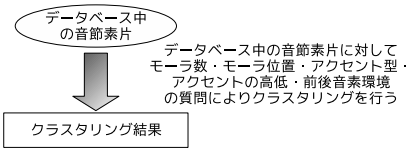

Fig. 2　標準モデル

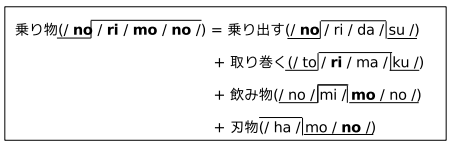

Fig. 3　標準モデルを用いた合成音声の例

*Word Synthesis by Concatenating Syllabic Components using Tree-based Clustering.　by UEMURA Kazuhisa,　MURAKAMI Jin’ichi and IKEHARA Satoru(Tottori University)

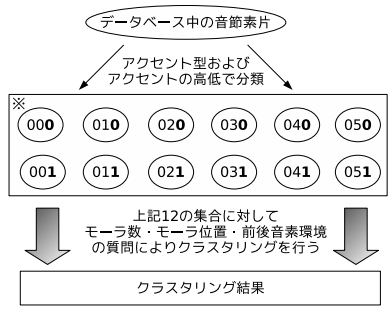

Fig. 4　拡張モデル

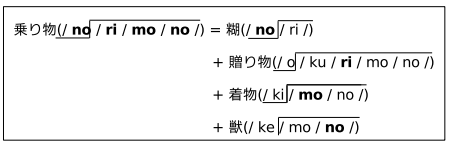

Fig. 5　拡張モデルを用いた合成音声の例

5　評価実験

5.1　実験条件

本章の実験では，まずデータベース中に存在する単語に対して，標準モデルと拡張モデルを用いて合成音声を作成し，音声品質の調査を行う．次にデータベースに存在しない単語に対しても同様の実験を行う．各モデルは HTK を利用して構築する．HMM の状態数は 1 とし，また HMM の学習データには，ATR 単語発話データベース Aset の 5,240 単語／話者の女性話者 1 名を用いる．作成する単語は，HMM の学習データに使用した自然音声，クラスタリングを使用せず音節波形接続型音声合成法で作成したオリジナル合成，標準モデルを用いて作成した音声，拡張モデルを用いて作成した音声の 4 種類とし，それぞれの音声に対して 3 または 4 モーラの単語各 100 単語を作成する．作成した単語の評価者は，音声研究に関わったことのない 5 名とし，評価方法にはオピニオン評価実験と対比較実験を用いる．また，各モデルにおけるクラスタリング結果の状態数は 1,000 程度になるように調整し，標準モデルの状態数を 1,000，拡張モデルの状態数を 998 とする．その他の実験条件に関しては，HTK の標準的なパラメータを使用する．

5.2　実験結果

データベース中に存在する単語のオピニオン評価実験と対比較実験の結果をそれぞれ Table2 および 3 に示す．またデータベース中に存在しない単語のオピニオン評価実験と対比較実験の結果をそれぞれ Table4 および 5 に示す．

Table 2 および 4 より，拡張モデルは標準モデルよりも高いオピニオンスコアが得られたが，両モデルに差が見られた．また Table 3 および 5 より，拡張モデルは標準モデルよりも良い音声であると分かる．

6　考察

6.1　実験結果に関する考察

データベース中に存在する単語に関する実験結果において，オピニオン評価実験では標準モデルと拡張モデルにあまり差が見られなかったが，対比較実験結果では大きな差が見られた (Table 2,3)．これは単語を単独で聞いた場合は，正しいアクセントでなくてもそれほど違和感を感じられないのに対し，同じ単語を同時に続けて聞いた場合には，アクセントの差違が明確になったのが原因だと考えられる．

データベース中に含まれていない単語に関する実験結果において，追比較実験ではデータベース中に存在する単語に関する実験結果と同様の傾向が見られたが，オピニオン評価結果では標準モデルと拡張モデルの音声品質に差が生じた (Table 4,5)．これは Table 4 の実験が，標準モデルと拡張モデルの単語のみで評価されたため，両モデルの差が明確になったことが原因と考えられる．以上より拡張モデルの有効性が示されたが，自然音声やオリジナル合成と比べると音声品質が悪いと考えている．

Table 2　データベース中に存在する単語のオピニオン評価実験結果 (総評価音節数：400)

音声の種類	オピニオンスコア
自然音声	4.50
オリジナル合成	3.83
標準モデル	2.51
拡張モデル	2.70

Table 3　データベース中に存在する単語の対比較実験結果 (総評価音節数:200)

	比較対象 1	比較対象 2
(1)	標準モデル：25%	拡張モデル：75%
(2)	自然音声：95%	拡張モデル：5%

Table 4　データベース中に存在しない単語のオピニオン評価実験結果 (総評価音節数：200)

音声の種類	オピニオンスコア
標準モデル	1.73
拡張モデル	2.63

Table 5　データベース中に存在しない単語の対比較実験結果 (総評価音節数:200)

比較対象 1	比較対象 2
標準モデル：24%	拡張モデル：76%

6.2　クラスタリングに関する考察

本論文で用いたクラスタリングは，音響的なパラメータを利用して行った．しかし自然音声やオリジナル合成と比べると音声品質の悪い音声であった．一方，オリジナル合成のように音響的なパラメータを使用せず言語的な情報のみで作成した合成音声は非常に良い音声品質が得られた．したがって音響的なパラメータを利用せず，言語的な情報のみを利用したクラスタリング手法を用いれば，音声品質が改善される可能性があると考えている．

7　おわりに

本研究では，クラスタリングを利用した音節波形接続型音声合成法の検討を行った．

データベースに存在する単語と存在しない単語をそれぞれ作成し，異なる 2 つのモデルに対して聴覚実験を行った．聴覚実験の結果，拡張モデルの有効性が示されたが，自然音声やオリジナル合成と比較すると音声品質の悪い音声であると分かった．今後は音響的なパラメータを使用せず言語的な情報を用いたクラスタリング手法を行う予定である．

謝辞　本論文を執筆するにあたり，参考にさせて頂いた論文，聴覚実験に協力してくださった研究室の方々に深く感謝いたします．

参考文献

[1] 村上 他. “音節波形接続方式による単語音声合成”, 信学論 D-II, Vol.J85-D-II, No.7, pp.1157-1165(2002)

[2] S. J. Young *et al.* “Tree-based state trying for high accuracy acoustic modelling. Proc”, ICASSP, pp.307-312(1994)

[3] “HTK Ver3.2 reference manual”, Cambridge University(2002)